Physics Applications Software at BNL

Torre Wenaus
Brookhaven National Laboratory
DOE HEP Scientific Computing Review
ANL
Feb 9, 2011

Outline

- Background: Physics Applications Software at BNL
- PAS contributors
- PAS in US ATLAS and ATLAS
- PAS Beyond ATLAS
- Looking to the Future
- Summary

Physics Applications Software at BNL

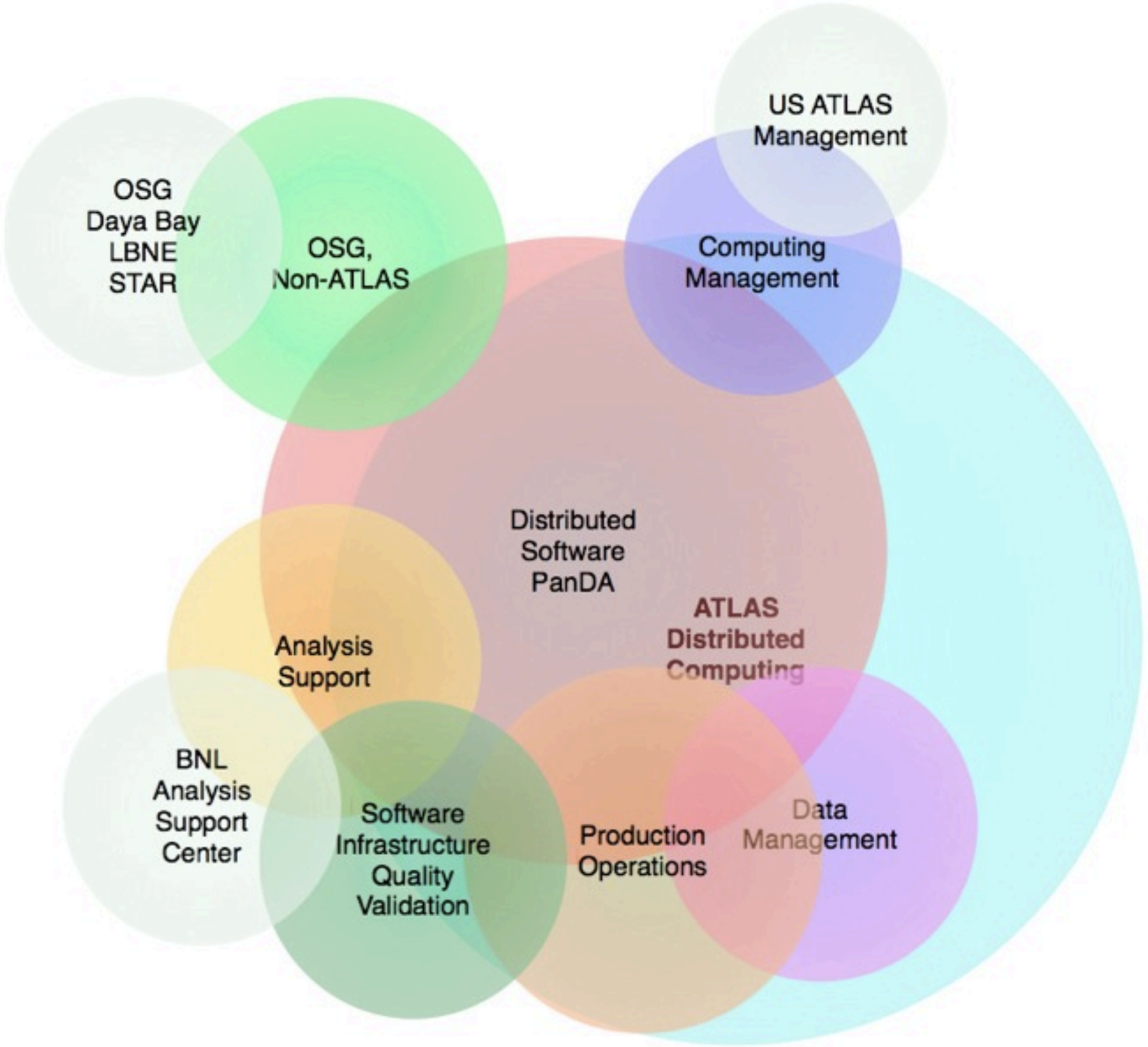
- Physics Applications Software (PAS) group in BNL Physics Department
  - Principal - but not sole - BNL home for this work
  - Founded a decade ago for primarily/initially ATLAS-directed PAS work in concert with BNL ATLAS physics program & BNL Tier 1, leveraging BNL expertise in the area
    - Many group members from STAR where they managed and developed the offline software system
  - Developed in concert with US ATLAS Physics Support and Computing project and its software priorities (next slide)
    - Original objective of broader application than ATLAS beginning to be realized
- Omega Group in BNL Physics Department
  - Home of the BNL ATLAS physics program and many HEP software experts among the group's physicists, drawing particularly on D0 experience
  - Original developers of ATLAS StoreGate transient event store (Rajagopalan, Ma)
  - Developers of the very successful D3PD Ntuple based analysis format (Snyder)
  - Host for US ATLAS Analysis Support Center (Ma), with one staffer (Ye) shared with PAS for applications software support
- Focus of this talk is the PAS group, as software professionals dedicated to work in this area

PAS Group Members

(C):Remote station at CERN (A):US ATLAS support (B):Base support (O):OSG support
(I): IT Computing Professional (S): Staff Scientist (P): HENP PhD
12 100% staff, 1 50/50 with Omega
Group Leader Alexei Klimentov

- David Adams (A,B,I,P) - Software quality & validation
- Jose Caballero (O,I,P) - PanDA@OSG, PanDA security
- Wensheng Deng (A,I,P) - Data management, analysis support
- Valeri Fine (A,I) - PanDA monitor
- Alexei Klimentov (C,A,I,P) - ATLAS Distributed Computing Coordinator, data management, production systems, operations
- Tadashi Maeno (C,A,I,P) - PanDA, analysis systems & support
- Pavel Nevski (C,A,I,P) - Production systems, operations
- Marcin Nowak (C,A,I) - Data management, PanDA DB
- Sergey Panitkin (A,I,P) - Analysis systems & support
- Maxim Potekhin (O,I,P) - PanDA@OSG, PanDA DB/monitor
- Alex Undrus (A,I,P) - Software librarian, software infrastructure
- Torre Wenaus (A,B,S,P) - US ATLAS Physics Support & Computing Manager, PanDA
- Shuwei Ye (A,I,P) - Software librarian, analysis support [joint with Omega]

PAS Activities & Interconnections



Size in very rough proportion to FTEs

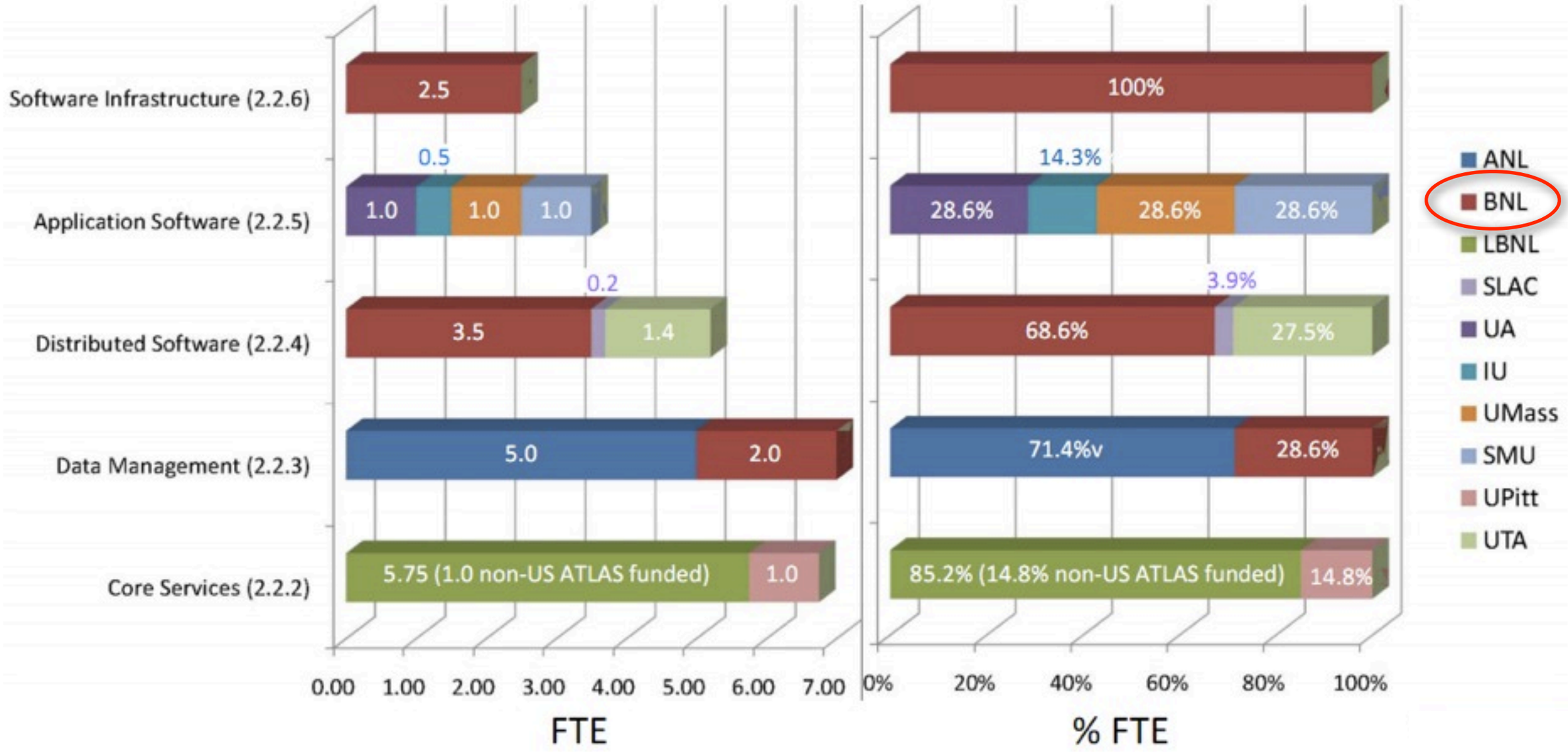
Grey bubbles are external entities

US ATLAS Physics Support and Computing Core Responsibilities

BNL PAS participation in bold

- Software
  - Core services: athena framework (centered at LBNL)
  - **Data management: event store** (centered at ANL)
  - **Distributed software: PanDA** (centered at BNL)
  - Detector-specific application software
  - **Software infrastructure support** (centered at BNL)
- Facilities and Distributed Computing
  - BNL Tier 1
  - Five Tier 2s across 9 institutions
  - Integrated U.S. Distributed Facility
  - Grid production: tools, services and operations
  - Tier 3 coordination
- Analysis Support
  - U.S. physics/performance forums
  - **Analysis tools, documentation and support centers**

US ATLAS Physics Support & Computing Activity/FTE Distribution



Peter Loch

US ATLAS Physics Support & Computing

BNL in bold

WBS Organization

2.1, 2.9 Management (Wenaus/Willocq)
2.2 Software (Loch; *Luehring from Feb 1)*
- 2.2.1 Coordination (Loch; *Luehring from Feb 1)*
- 2.2.2 Core Services (Calafiura)
- 2.2.3 Data Management (Malon)
- **2.2.4 Distributed Software (Wenaus)**
- 2.2.5 Application Software (Luehring; *Neubauer from Feb 1*)
- **2.2.6 Infrastructure Support (Undrus)**
- *2.2.7 Analysis support (retired; redundant)*
- 2.2.8 Multicore Processing (Calafiura)

2.3 Facilities and Distributed Computing (Ernst)
- **2.3.1 Tier 1 Facilities (Ernst)**
- 2.3.2 Tier 2 Facilities (Gardner)
- 2.3.3 Wide Area Network (McKee)
- 2.3.4 Grid Tools and Services (Gardner)
- 2.3.5 Grid Production (De)
- 2.3.6 Facility Integration (Gardner)
- 2.3.7 Tier 3 Coordination (Yoshida/Benjamin)

2.4 Analysis Support (Cochran/Yoshida)
- 2.4.1 Physics/Performance Forums (Black)
- 2.4.2 Analysis Tools (Cranmer)
- **2.4.3 Analysis Support Centers (Ma)**
- 2.4.4 Documentation (Luehring)

ATLAS Distributed Computing (ADC)

US
BNL

Distributed Computing
A.Klimentov

Software Development
G.Stewart

Operations
S.Jezequel

ADC ACTIVITIES

Information Protection
G.Poulard

Documentation
M.Smizanska

Tier-3 Technical Coordination
D.Benjamin

Central Services& Security
F.Donno

Grid Software Releases
A.DeSalvo
L.Rinaldi (deputy)

CERN Data Processing and Analysis
A.DiGirolamo

Tier-0
L.Goossens
A.Nairz

Databases
D.Barberis

Distributed Data Management
S.Campana

Distributed Analysis
J.Elmsheuser
D. van der Ster

ADC-CERN
J.Schovancova
J.Yu

Grid Data Processing
K.De
X.Espinal
P.Nevski

DB Ops
D.Barberis

DDM Ops
I.Ueda

DAST
N.Ozturk
D. van der Ster

AMOD

ADCoS
J.Schovancova

FroNTier/squid
J.DeStefano
R.Walker

DA Tools
T.Maeno

COMP@P1
G.Negri
J.Schovancova
J.Yu

PanDA infrastructue
D.Smith

SQUADS

ADC PRODUCTS

AGIS
A.Anisenkov
A.Klimentov

AMI
S.Albrand

DQ2 & Data Management
S.Campana

GANGA
J.Elmsheuser

Monitoring
A.Read

PanDA
T.Wenaus

ProdSys
D.Smith
G.Stewart

A.Klimentov . v0.8 Sep 25 2010

All positions are one year appointment and they are rotational

PAS in ATLAS Distributed Computing

The largest PAS activity area - 10 PAS members involved

- ADC project leadership (Klimentov)
- PanDA distributed production and analysis system leadership and principal developer roles (Wenaus, Maeno)
- Production systems development and operations (Nevski, Klimentov)
- Production quality assurance and validation (Nevski)
- Production data distribution systems and operations (Klimentov, Nevski, Deng)
- Analysis systems development and user support (Maeno, Panitkin, Deng)
- PanDA/production system database development and support (Potekhin, Fine, Nowak, Wenaus)
- PanDA monitoring systems (Fine, Wenaus)
- ATLAS grid information system (AGIS) (Klimentov)
- PanDA systems analysis and user behavior (Panitkin)
- PanDA system security (Caballero)

PanDA Distributed Production and Analysis System

The largest PAS software effort - 8 PAS members involved in PanDA & related (4 in core US ATLAS supported PanDA effort, 3.5 FTEs)

- Initiated in 2005 as the US component of the ATLAS production and analysis system
- Design grounded in scalable, proven, off-the-shelf web and middleware technologies (apache+python, relational DBs, http messaging, Condor)
- Based on 'pilot jobs' to minimize grid/site failure modes, maximize flexibility and performance in job management and brokering
- Global job queue for simple systems view and brokerage
- Tightly integrated with data management, data flow
- Success drove ATLAS-wide adoption in 2008
- Has been accruing further responsibilities since
  - eg. in growing data management role
  - approaching universal adoption as the ATLAS distributed analysis system (usage of the 'other' system Ganga is declining, <20%)

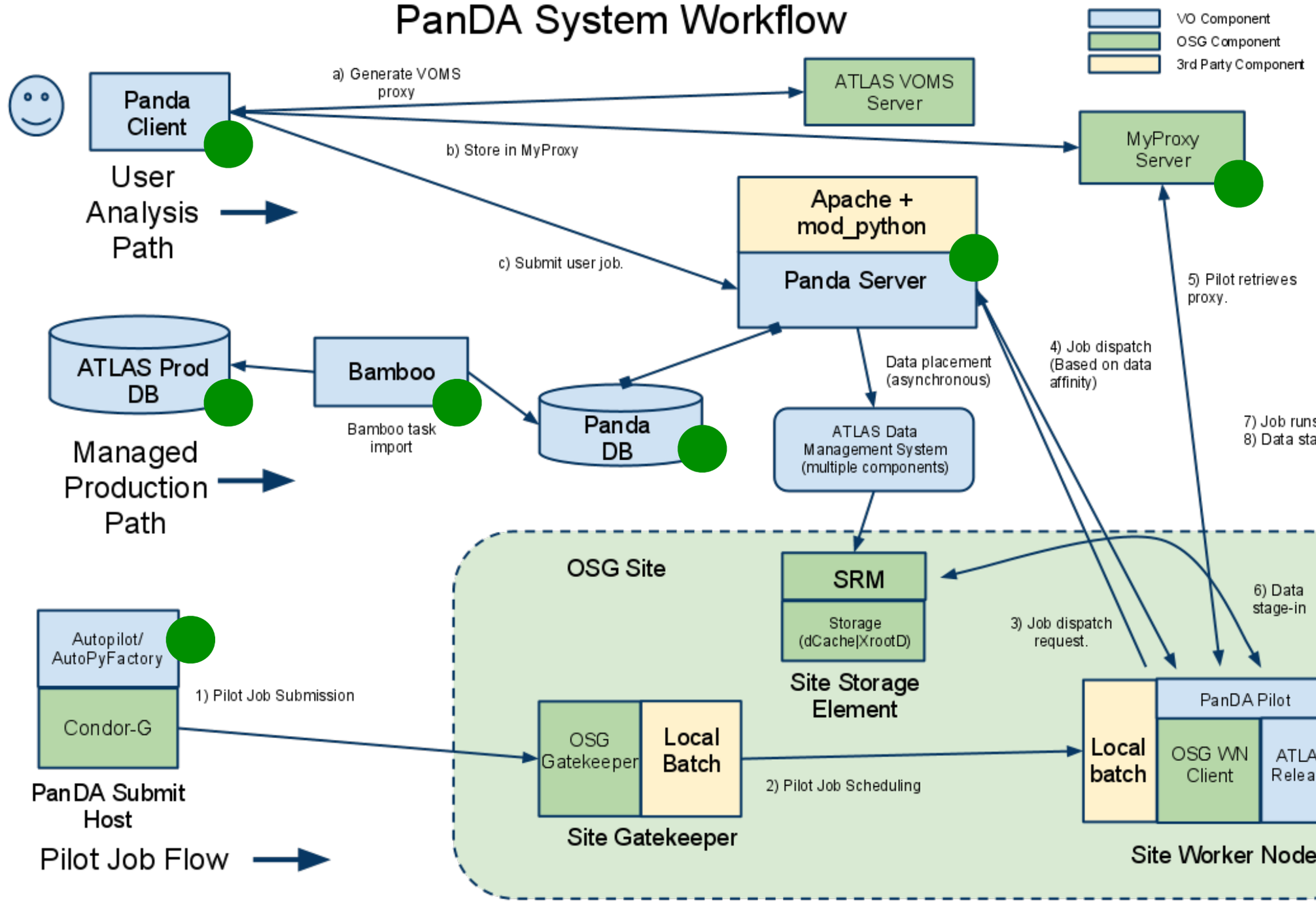
PanDA Functional Areas

BNL responsibilities in bold

- **PanDA server - the system core services**
  - Job queue management, associated data management, job brokerage, job dispatch to pilots
- **Bamboo - PanDA interface to ATLAS production DB**
- PanDA pilot - requests payload job and manages job execution environment (UT Arlington)
  - In principle a simple function; in practice, complex
  - Much of the heterogeneity of the grid, particularly in storage services, is encapsulated by the pilot
- **PanDA monitor - operational interface for production operators, analysis users**
- **Pilot factory - management of pilot submission to sites**
  - Responsibility recently moved out of PAS to BNL Tier 1
- Site configuration DB - site configuration and control parameters (UT Arlington)
- **PanDA-based analysis - pathena and prun**
  - Full-functioned (and continually evolving) analysis front-end to PanDA for athena (pathena) and generic - typically ROOT - user jobs (prun)
- **PanDA-based data management - PandaMover, PanDA dynamic data placement (PD2P)**
  - PanDA used as an intelligent workload-aware driver for data movement, utilizing ATLAS data management tools
- **PanDA system security - glexec identity-switching service and associated infrastructure**

PanDA Schematic

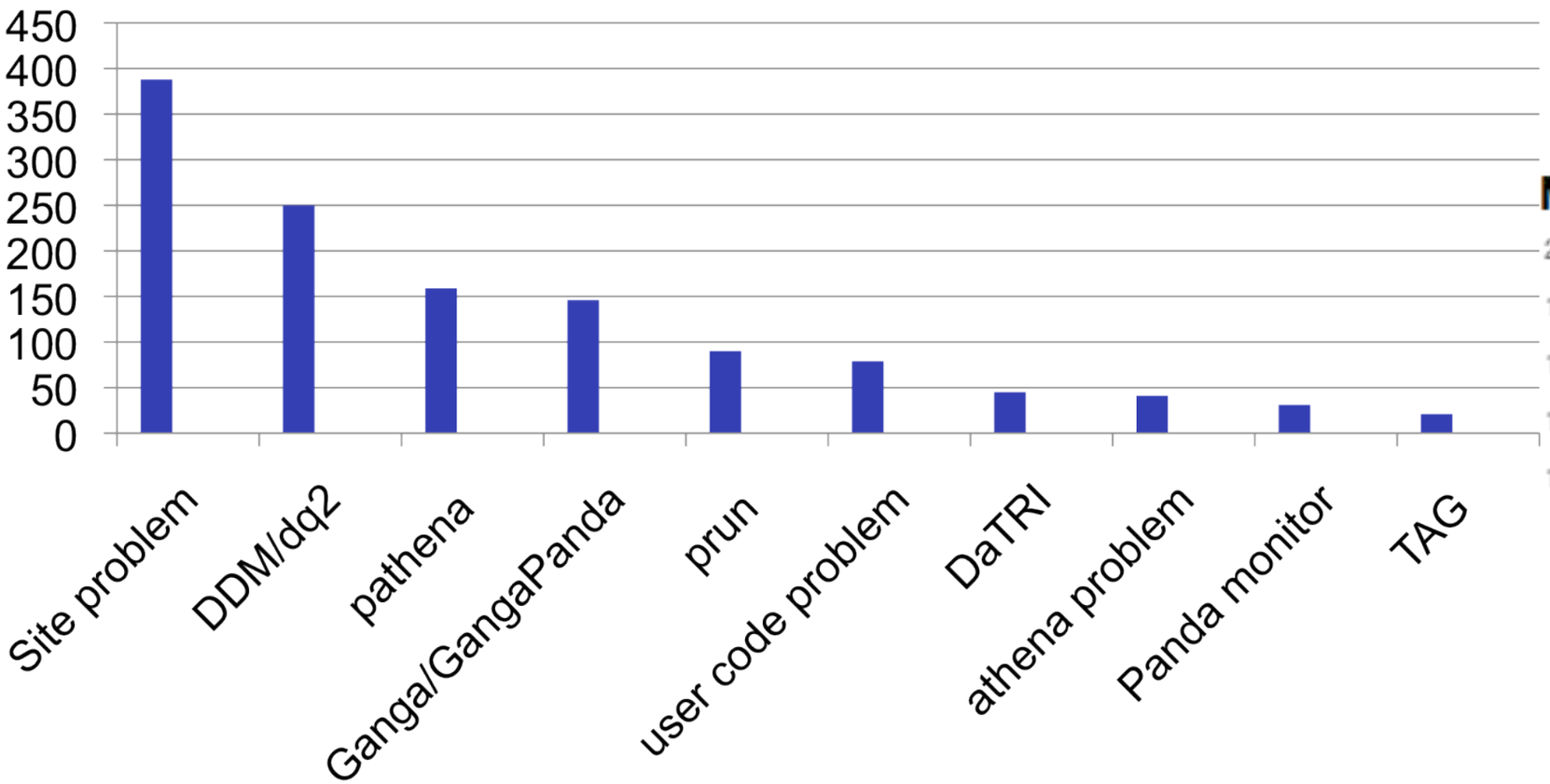
- Major BNL role

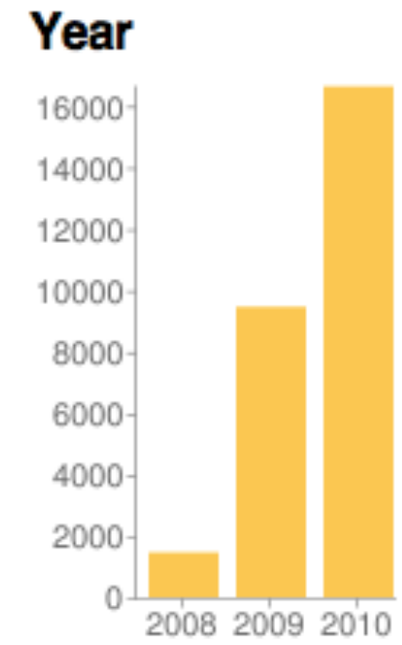


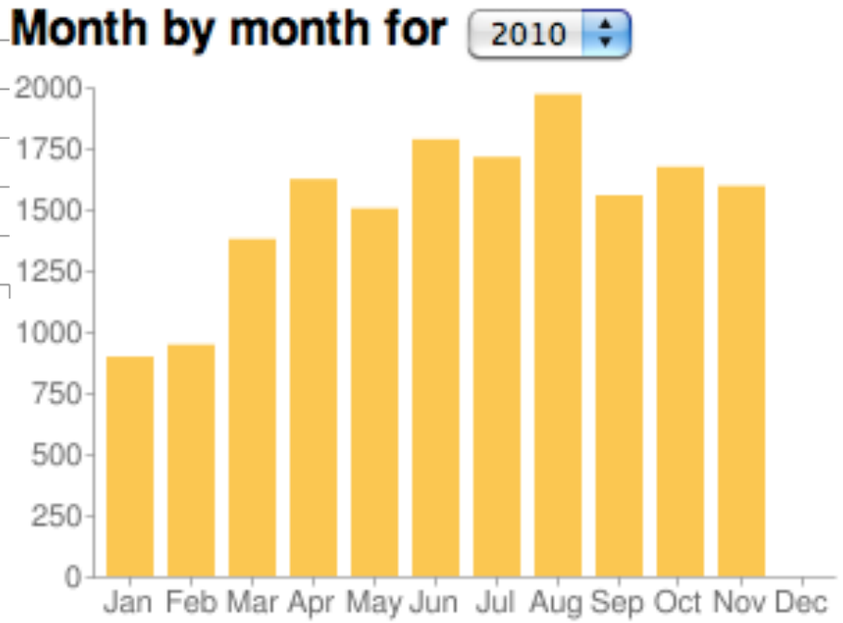
John Hover

Distributed Analysis Support Team (DAST)

- Direct support for analysis users via shifts responding to user analysis questions and problems
- BNL a major contributor: Deng, Panitkin, Ye (3 of of the 8 DAST members in US zone)
- ~6000 threads, 27500 messages 10/08-11/10





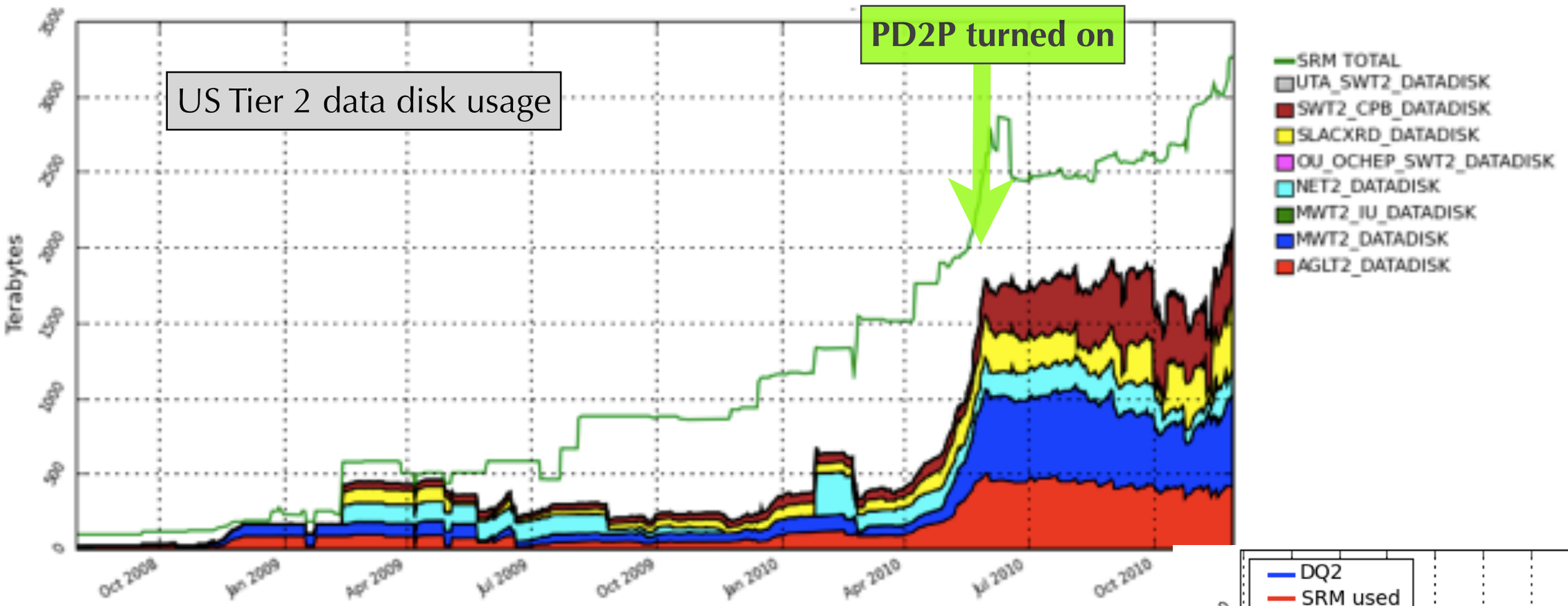


Nurcan Ozturk

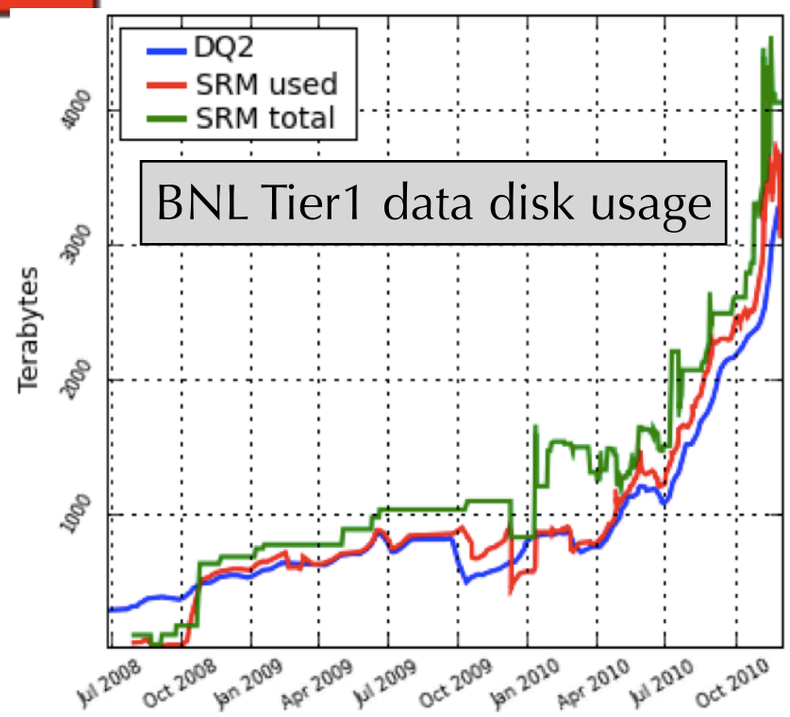
Key Issues for ATLAS Computing in 2011

- As US ATLAS PS&C recently told our reviewers:
- Principal issue for ATLAS computing in 2011 is surviving the data onslaught! Up to 100x integrated luminosity, high (400Hz) trigger rate, only ~30% more computing resources than 2010
- PAS making key contributions to this
  - PanDA extensions dynamic data management extensions in 2010 critical to not exhausting our disk space; extending for 2011
  - Studies of system performance and user behavior (eg. in chosen data formats) driving decision making and system optimization
  - Major program in addressing DB scalability through ramped Oracle involvement and prototyping 'noSQL' databases (Cassandra, SimpleDB, ...) as potential (partial) alternative
  - Studying application of WAN based, fine-grained (event level) data access and caching

PanDA Dynamic Data Placement (PD2P)

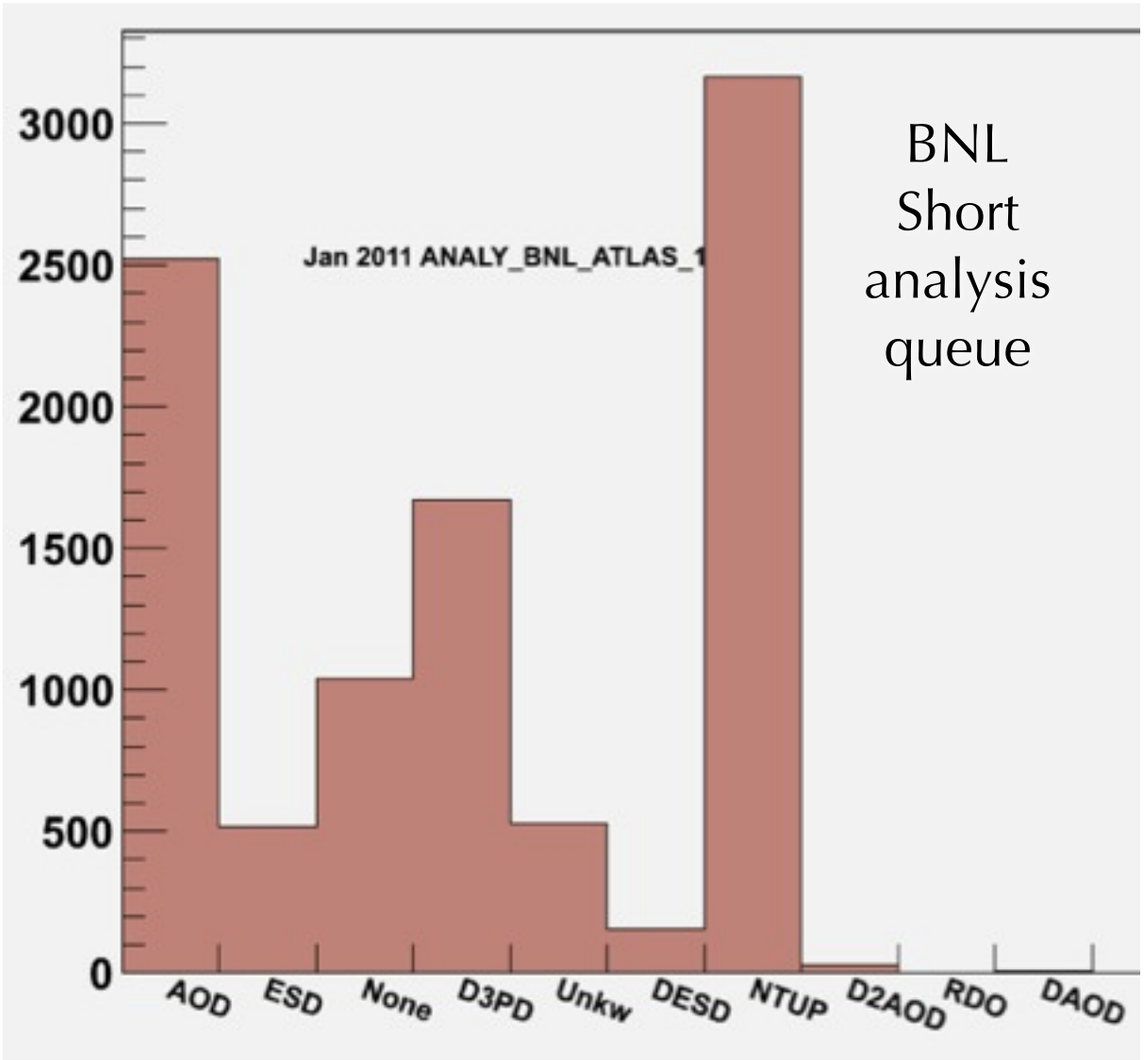


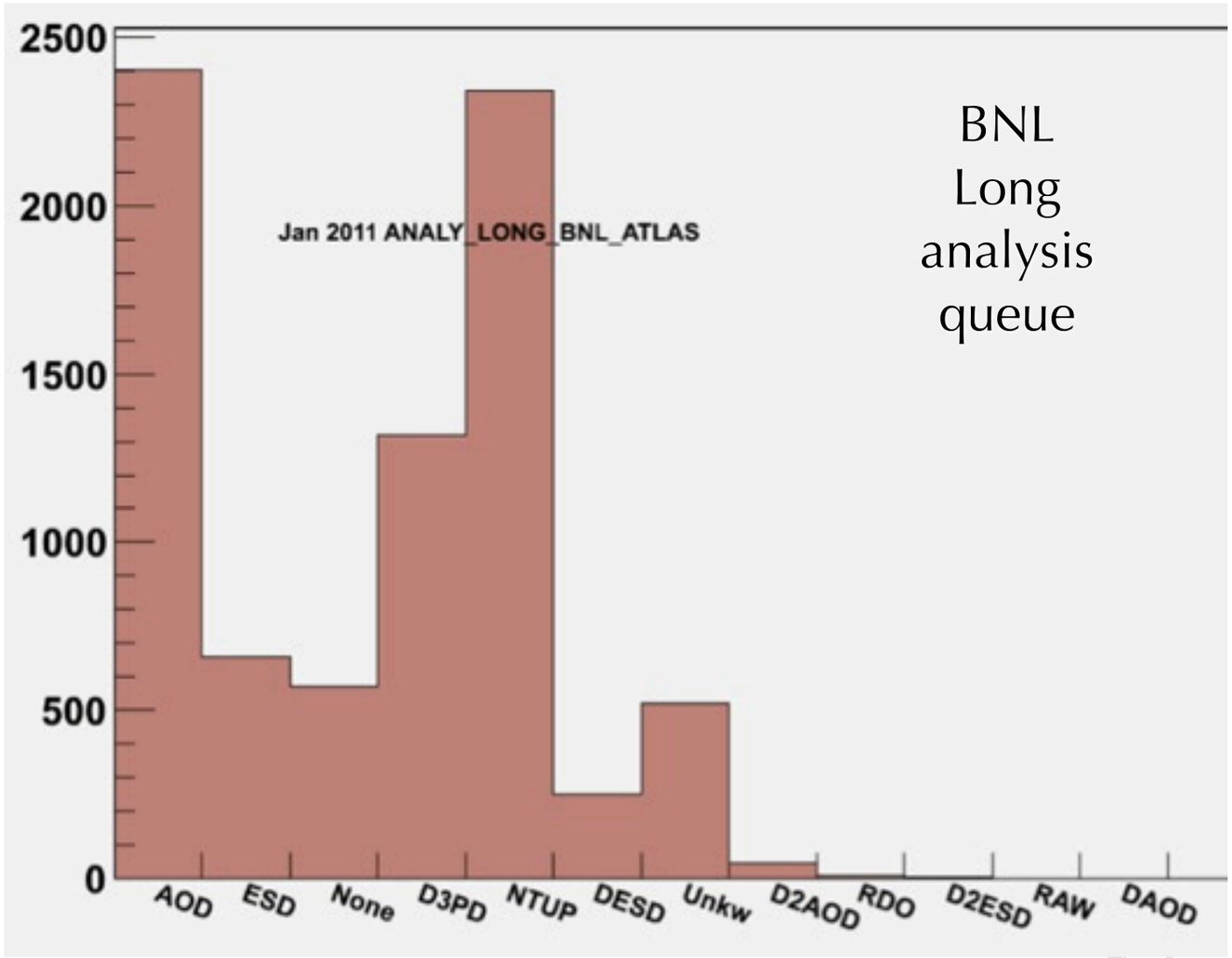
- PanDA's PD2P, introduced 6/10, sends data to Tier 2s for analysis use on the basis of usage, rather than pre-defined policy based distribution
- Plot above shows effect of turning it on in June in the US
- Flattened exponentially rising consumption, addressing an otherwise critical space usage problem
- Since extended to T2 space mgmt in all ATLAS clouds
- Next issue: Tier 1s – plan to extend PD2P there for 2011 datataking



Analysis System Usage/Performance Studies

- Sergey Panitkin uses Panda archival logs to study user behavior and system performance for Panda analysis
- Latest studies: addressing key question (for space usage) of data popularity measures - are users migrating from expensive/bulky formats (ESD) to space-efficient formats (AOD, D3PD). Answer is yes - input to decision making on viability of dropping most ESD storage





Large Scale Disk Pool Demonstrator 'LST2010'

- A large scale storage testbed for real-world testing
  - O(1PB) storage, O(1k) cores
  - Test data handling scenarios using xrootd-based disk pool plus functional extensions from CERN IT-DSS
  - Use the testbed for ATLAS distributed analysis and production, testing data handling/caching scenarios
- Negotiated and managed by Klimentov as ADC/CERN IT testbed
- Integrated in PanDA/PD2P by Maeno
- Successfully deployed as operational PanDA analysis site in 2010, ~500 job slots
- Working to expand the scale in 2011
- BNL Tier 1 preparing a similar test with a different back end storage technology (BlueArc) and larger job slot count

Beyond PD2P: Event Level Caching

- New effort integrated with & building on the Large Scale Disk Pool Demonstrator work
- PD2P makes data movement dynamic and usage-driven but the distributed data management (DDM) scheme is fundamentally the same: replicating datasets to sites
- Motivations to move beyond PanDA/DDM-based dynamic data placement to a caching system:
  - **DDM simplification**: data movement is inherent in the caching system, not explicitly driven by DDM
  - **Finer granularity**: file or page level
    - More efficient use of space & network
    - Better utilization of sites with little space
  - **On demand:** data is moved when it is needed, transparently
- New ROOT/ATLAS developments make this possible
  - ROOT TTreeCache + xrootd for sub-file level caching
  - ROOT and CMS work have demonstrated its viability
  - Panitkin will work with PanDA, event persistency, facility and CERN IT teams to try this out in 2011
  - Maeno will extend PanDA data-aware brokerage to provide a 'cache memory' to send jobs where the needed data has been previously cached

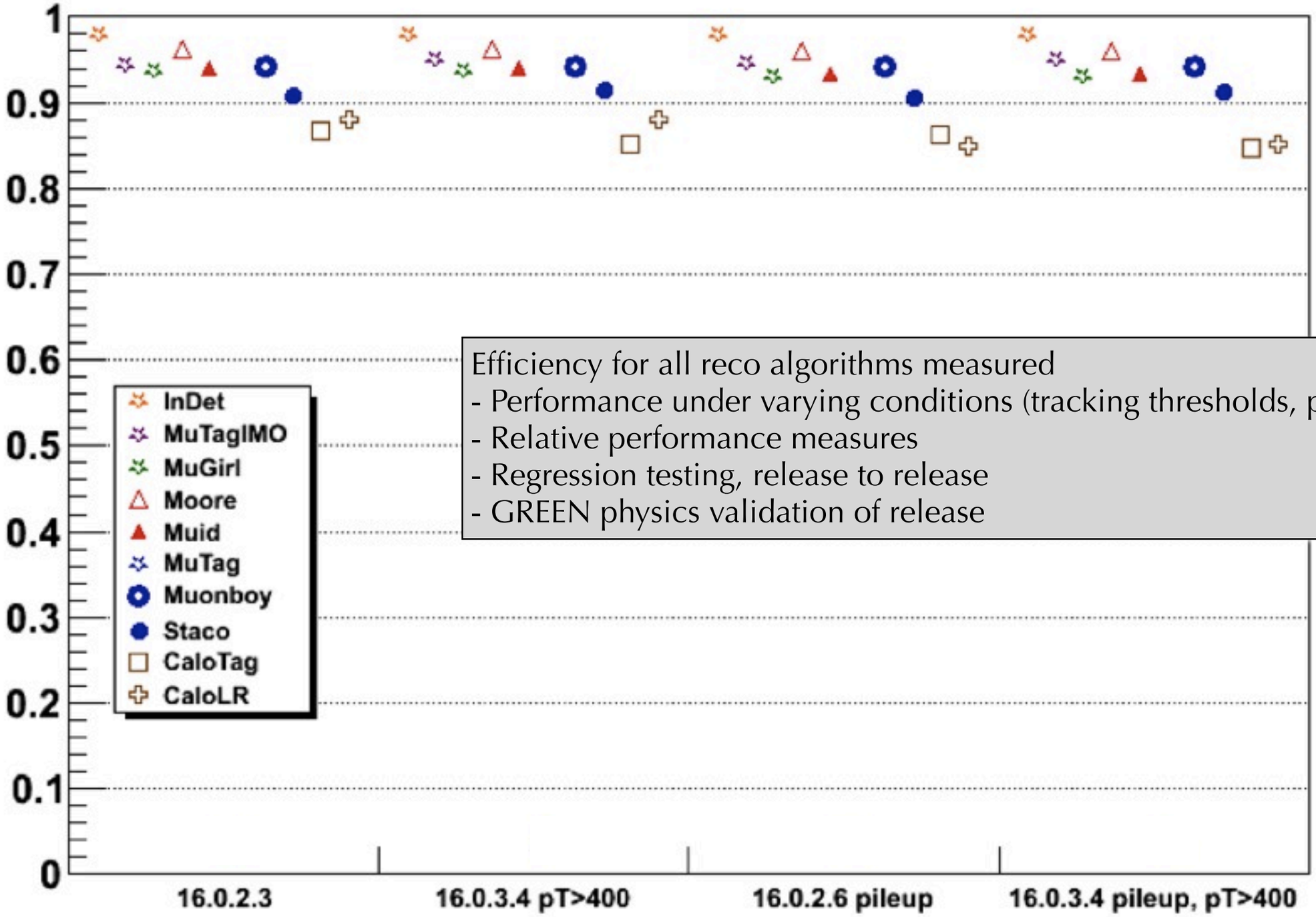
Other ATLAS Computing Roles

- ATLAS event persistency
  - Marcin Nowak is responsible for the interface between underlying POOL/ROOT persistency services and the event tag system for fast event selection in analysis
    - Works as an integral member of the ATLAS event persistency team centered at ANL
  - We also draw on his Oracle expertise for PanDA DB work
- ATLAS software infrastructure
  - Alex Undrus is responsible for the ATLAS nightly software build and test systems underlying developer and release support
- ATLAS software quality and validation
  - David Adams is responsible for software quality and validation for combined muon reconstruction, complementing his (base supported) physics work
- US ATLAS Analysis Support Center (ASC) @ BNL
  - Shuwei Ye provides software and analysis systems support to US ATLAS analysis users via the ASC, complementing his software infrastructure and distributed analysis support roles

ATLAS Nightly Build System (NICOS)

Alex Undrus

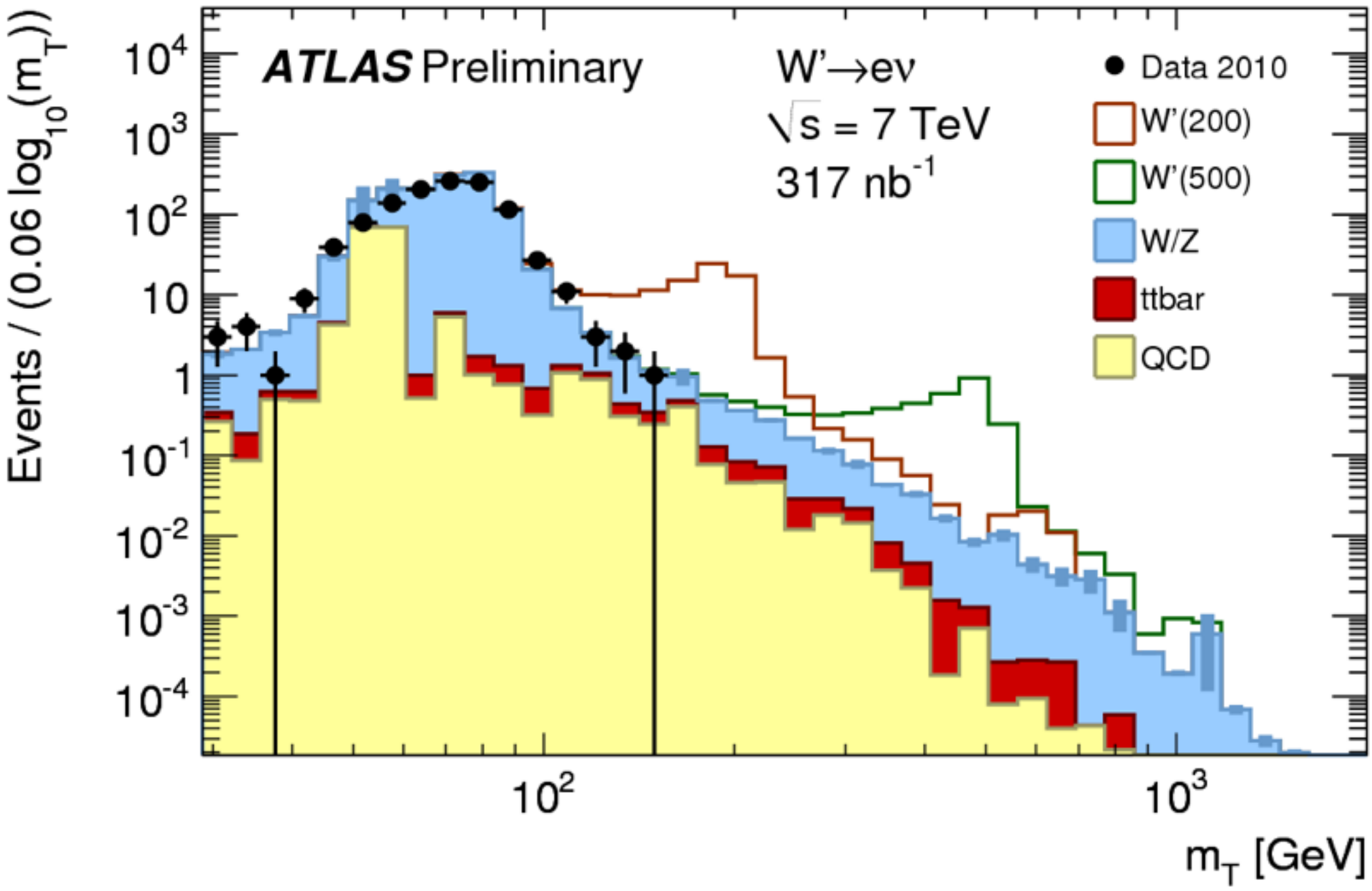
Combined Muon Reconstruction Performance



Efficiency for all reco algorithms measured
- Performance under varying conditions (tracking thresholds, pileup,..)
- Relative performance measures
- Regression testing, release to release
- GREEN physics validation of release

Physics Analysis Participation

David Adams (partially base supported) complements his muon reco QA/validation work with muon/electron driven physics studies such as this W′ search



David Adams

PAS Beyond ATLAS

- Original objective of PAS group to support BNL (and wider) HENP software/computing beyond ATLAS is beginning to be realized
- PanDA supported as an OSG workload management system
  - Long-time user group in structural biology
  - Two PAS staff (Caballero, Potekhin) supported by OSG extensions program in workload management
- PAS working with Daya Bay/LBNE collaborators at BNL to apply PanDA as distributed production system (Caballero, Potekhin)
  - PanDA for Daya Bay now operational at PDSF (LBNL)
- PanDA ported to STAR as a offline production prototype,

Looking to the Future

- ATLAS computing R&D areas: 'active' = active in US; '**active**' = active at BNL
- Collaborative, seeking more/broader collaboration (eg. OSG)
  - Virtualization and cloud computing **(active)**
  - Multi/many-core computing (active, supplemental DOE support)
  - Campus grids and inter-campus bridging (active)
    - Bring to the campus what has worked so well over the wide area in OSG
  - 'Intelligent' cache-based distributed storage **(active)**
    - Efficient use of disk through greater reliance on network, federated xrootd
  - Hierarchical storage incorporating SSD **(active)**
  - Highly scalable 'noSQL' databases **(active)**
    - Tools from the 'cloud giants': Cassandra, HBASE/Hadoop, SimpleDB…
  - 'Flatter' point-to-point networking model (active)
    - Validation, diagnostics, monitoring of (especially) T2-T2 networking
  - GPU computing (active in ATLAS, not in US)
  - Managing complexity in distributed computing **(active)**
    - Monitoring, diagnostics, error management, automation

Summary

- HEP Physics Applications Software as a dedicated activity at BNL has grown up with ATLAS computing at BNL and in the US, and has drawn heavily on BNL's software expertise from STAR, D0
- Strong complementary roles within US ATLAS PS&C program
- Well integrated with lead roles in ATLAS computing
- Dominant role (at ATLAS as well as US level) in ATLAS distributed computing and its production, analysis and data management systems
- Accrued expertise can be applied elsewhere and we are pursuing this: OSG, BNL neutrino physics program, ...
- Collaboration on many axes on present and planned activities: (US) ATLAS colleagues, Facilities, OSG, CERN IT
- Central roles in ATLAS scale-up, and forward-looking R&D activities to keep abreast of the scaling and technology curves